科学知识网络扩散中的社区扩张与收敛模式特征分析*
——以医疗健康信息领域为例

■ 岳丽欣[1] 周晓英[1] 刘自强[2,3]
[1] 中国人民大学信息资源管理学院 北京 100872 [2] 中国科学院成都文献情报中心 成都 610041
[3] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190

摘 要：[目的/意义]科学知识网络中知识单元呈现出一定的集群性与社区性，揭示科学知识网络扩散时序变化过程中的社区扩张与收敛的基本模式与特征，对于拓展、深化科学知识扩散与传递规律研究具有一定的意义。[方法/过程]首先，基于引用关系建立邻接矩阵进而构建学科知识网络，采用复杂网络分析中的 Louvain 社区探测算法对领域知识网络进行社区划分；然后利用网络表示学习技术进行社区扩张与收敛特征表示与计算；最后以时间序列为逻辑线索，对不同社区的扩张、收敛演变过程进行动态跟踪建模，从而揭示科学知识网络时序变化过程中社区扩张与收敛的基本模式与特征。[结果/结论]以医疗健康信息领域进行案例研究，研究发现社区扩张模式的发展趋势符合 S 形曲线函数中的 Logistic 模型，社区收敛模式的发展趋势符合 S 形曲线函数中的 BiHill 模型。

关键词：知识网络 社区探测 网络表示学习 扩张模式 收敛模式
分类号：G251
DOI：10.13266/j.issn.0252-3116.2020.14.007

1 引言

在数据科学时代背景下，随着科技创新环境不断完善、科学文献数量爆发式增长、数字出版行业及其相关技术的快速发展，推动全球范围内的科学文献传播，加速了世界科学知识交流，进一步促进了科学技术的协同发展。目前企业界、学术界十分重视对海量科学文献的存储、挖掘和利用等，但是关于科学文献中蕴含的知识增长、传播影响因素、基本特征和机理规律等问题的研究有待深化。

在科学文献传播过程中，文献作为知识传播的主要载体，通过作者的引用、合作等关联关系形成纽带构成科学知识网络，随着时间的推移，科学文献及其关联不断增多，承载的知识不断传递与扩散。探索科学文献网络时序变化的基本过程，揭示科学知识传播规律，为科学技术决策提供参考依据，对促进科技创新具有重要的意义。近年来，科学知识网络的扩散、演化相关研究成为图情领域学者关注的重点之一，其中“知识节点”由科学文献或者相关概念、事物（知识要素、知识单元等）表征，“边”通过“知识节点”之间的某种关联关系（共现、引用和因果等关系）进行联系。概括来讲，众多学者以定量化、知识化和网络化的视角探索客观科学知识网络，通过分析科学知识网络从而研究揭示知识扩散、知识传递等现象的客观规律。

目前，学者的研究主要通过分析微观视角下科学文献之间的引用关系探索知识扩散，但由于知识单元都不是以完全孤立与游离的状态存在，而是基于关联关系呈现出一定的团簇性与集群性（社区），通过微视角下的科学文献引用关系探索知识扩散（知识扩散速度、广度等），便无法有效揭示知识扩散过程中的集群（社区）时序变化机理，所以对于集群（社区）维度下的知识扩散研究有待深化。笔者尝试探索并揭示科学知识网络扩散过程中，引文网络社区的扩张与收敛模式与特征，从而拓展、深化科学知识网络的扩散、传递相关研究。

* 本文系国家自然科学基金项目“医疗健康网站信息可信度与质量控制研究”（项目编号：71473260）和国家社会科学基金项目“健康中国建设中的国民健康促进和健康服务策略研究”（项目编号 16AZD021）研究成果之一。
作者简介：岳丽欣（ORCID：0000-0002-7268-7871），博士研究生；周晓英（ORCID：0000-0002-9116-1525），教授，博士生导师，通讯作者，E-mail：xyz-ruc@qq.com；刘自强（ORCID：0000-0003-1814-8655），博士研究生。
收稿日期：2020-02-12 修回日期：2020-04-23 本文起止页码：63-73 本文责任编辑：徐健

2 文献综述

2.1 科学知识扩散

目前研究者将引文网络和合作网络作为知识扩散研究的主要载体,以定量化、网络化和知识化的视角探索客观科学知识网络扩散,揭示显隐性知识的扩散、传递等现象的客观规律,产生了大量优秀研究成果。

邱均平等[1](2014)在构建国内知识图谱领域引文网络的基础上,从网络中文献的期刊、机构、作者、关键词 4 个层面进行整合、细化引文网络,并引入时间维度进行分析以揭示知识的扩散与演进过程,发现国内知识图谱研究由科技管理领域扩散到图书情报领域,进而推广应用于教育学等学科领域。李纲等[2](2017)引入复杂网络和超图数学理论构建了一种基于科研合作超网络的知识扩散演化模型,以探究科研合作网络中知识扩散的演化规律和动力机制;并且通过再现真实的网络组织知识传播行为,揭示了科研合作网络中不同网络结构特征、结点偏好性选择、知识增长老化以及知识扩散途径与知识传播扩散过程的动态关系。岳增慧等[3](2019)以社会网络领域为例分析了学科引证网络知识扩散特征,具体以文献引证作为学科知识传播路径载体,从集中趋势、离散程度和分布形态 3 方面对学科知识扩散中介性特征以及中间人角色特征进行剖析,结果表明社会网络领域学科间的知识交流活动频繁,学科知识扩散数量特征指数波动范围较大,离散程度较高,分布多呈现长尾偏右的尖顶曲线。科学知识扩散基本规律研究一直是图书馆学、情报学学者关注的重要问题之一,但是目前的研究侧重于通过微视角下的科学文献引用关系探索知识扩散(知识扩散速度、广度等),知识扩散过程中的集群(社区)时序变化机理相关研究有待深化。

2.2 科学知识增长自组织机制

科学知识增长问题一直是情报学、科学学等学科领域学者研究的重要问题之一,目前研究结果表明,随着时间的推移各个学科领域的科学知识增长具备自组织特性。

其中代表性成果主要有,E. C. M. Noyons 等[4](1998)在研究中指出科学技术发展过程中,科学认知系统通过科学知识的自组织推动科学知识的动态增长与老化。此外,关于科学知识增长的本质问题,有的学者从社会学术交流角度进行探索,如 L. Leydesdorff 等[5](2003)认为科学知识增长的本质是一个科学社会内部的自组织过程,在具体研究中将整个科学界所有学科看作是一个大系统,而每一学科看作是整个科学大系统中的子系统[6-7],借鉴系统论动力学的思想来研究科学知识的动态增长规律,通过研究表明,科学研究活动通过社会化的学术交流来获得知识的积累与增长;有的学者从科学哲学角度探讨科学增长问题,如 K. Popper[8](2007)在古典的经验主义、理性主义和批判理性主义等方法论基础上,指出科学始于问题,科学理论就是对科学问题的试探性答复,科学发展是一个从问题到问题的链式过程,通过对先前知识的修改而获得新的知识,促进科学知识增长,这种“链式”结构使得科学知识增长具备自组织特性;还有的学者从逻辑学角度探讨科学知识增长问题,如靖继鹏、马费成等[9](2009)认为科学是一个不稳定的逻辑混乱系统中通过逻辑合理化组织来获得的稳定有序性建构,通过知识的自组织促进科学系统的有序化,推动科学知识的增长。在情报学领域中,学者主要通过科学文献研究科学知识增长问题,如刘则渊等[10](2012)基于科技文献数据,利用共被引分析、共词分析和可视化分析等文献计量学方法来分析科学知识结构的动态演化过程,并在研究中指出科学发展过程中会发生科学知识单元的分解和会聚、离散和重组、演进和升华、衍生和转化,形成一个从简单到复杂、从低级到高级、从混沌到有序的自组织系统;万昊[11](2017)认为随着系统论的思想被引入科学计量学领域,科学知识增长问题在系统论的框架有了新的阐述,即科学知识增长过程中通过逻辑合理化组织以获得稳定有序的网络结构。

科学知识增长的自组织特性在微观层级上导致离散知识单元析出、游离并重组、更新,进而形成知识单元的科学知识网络结构。由于知识单元都不是以完全孤立与游离的状态存在,而是基于关联关系呈现出一定的团簇性与集群性(社区),科学知识网络扩散中伴随着社区结构的扩张与收敛现象。

2.3 知识网络社区演化

学科领域内部知识单元都不是以完全孤立与游离的状态存在,而是基于显隐性关联关系呈现出一定的团簇性与集群性[12],一些学者尝试探索科学知识生长过程中的知识团簇性与集群性演化机理[13]。

其中,复杂网络社区这一概念的提出有效推动了知识网络社区演化有关研究。M. Girvan 等[14](2002)提出了社区结构(community structure)概念,把社区界定为复杂网络中的一个子图、子网络,基本特征为社区内部节点间链接紧密,不同社区间链接稀疏,并指出网络社区(network community)或网络团簇结构(network

cluster)是复杂网络最普遍和最重要的拓扑结构属性之一,在此基础上图情领域学者对复杂网络社区演化展开了大量研究。比如,M. A. Bettencourt 等[15](2009)基于网络密度、直径、连通性等指标揭示知识网络社区的时序演化过程,研究中指出一个领域从产生到成熟,可以理解为作者们在初期进行离散的孤立的研究,之后逐渐交融形成统一认识的过程;白如江[16](2013)、王晓光[17](2013)等通过分析知识网络社区的演化情况,挖掘、揭示相关领域研究主题的发展趋势;滕广青[18－19](2018)以知识间关联关系为基础,从频度、关联、数量、规模、时序多个维度进行交叉复现分析,对社会化标注模式下 Folksonomy 知识组织模式中领域知识群落的生长展开研究,研究指出知识群落生长模式与规律的揭示,有助于从知识间的互促互扰关系方面拓展领域知识组织视野并把握知识发展的脉络。

知识网络社区演化相关研究是复杂网络思维下对科学知识生长规律研究的拓展、深化,特别是数据科学时代背景下科学文献爆发式增长,科学知识联系愈发紧密,针对知识网络结构关系(网络社区,团簇性与集群性)的时间序列分析,更有助于揭示知识发展演化过程中交叉、衍生、融合等现象背后的规律。

2.4 科学知识在科学共同体中的扩散

T. S. Kuhn 在《科学革命的结构》一书中提出科学发展模式理论,以范式为核心概念,认为科学发展是一个科学范式不断变迁的过程并永无止境不断发展,并将科学范式[20]定义为“某一学科领域共同体的共同约定”。

D. J. Price 以“无形学院”这一概念来指那些从正式的学术组织中派生出来的非正式学术群体(科学共同体),研究中探索科学内部的社会结构与科学知识增长的关系,即学科或专业的社会组织与知识增长的关系,并以学术期刊为例对 1650 年－1950 年的学术期刊数量增长规律进行研究,提出了文献指数增长规律[21－22]。D. Crane 在 T. S. Kuhn 的科学发展范式理论、科学共同体学说和 D. J. Price 的无形学院、科学知识增长定量研究基础上,通过学术论文引证关系分析了科学家之间的联系,目的是通过具体可察的数据资料来说明科学家之间的种种非正式的不固定的社会联系以及各个学科领域中的“无形学院”的存在,并分析了科学知识和科学共同体在科学文献增长曲线不同阶段的特征[23],具体内容见表 1。

表 1　科学知识和科学共同体在不同文献增长阶段的特征

名称	阶段 1	阶段 2	阶段 3	阶段 4
知识特征	范式出现	常规科学	重大问题的解决	衰竭
			反复出现	危机
科学共同体特征	社会组织少	合作者群体	日益专业化	成员减少
	或无	或无形学院	争论日益加剧	成员减少

由表 1 分析可知,科学知识增长、扩散过程中与科学共同体(社区、社群、群落)密切相关。在阶段 1 中,新范式出现吸引部分科学家,科学共同体(社区、社群、群落)特征不明显;在阶段 2 中,随着研究的深入,科学知识增长、扩散吸引大量科学家形成科学共同体(社区、社群、群落);在阶段 3 中,随着研究愈发深入,呈现日益专业化特征,科学共同体(社区、社群、群落)趋于稳定;在阶段 4 中,由于科学的发展,部分研究衰退,科学共同体(社区、社群、群落)成员减少,科学知识发展衰退特征显著。D. Crane 关于知识在科学共同体中扩散的研究,对于目前科学知识扩散有关研究具有理论借鉴意义;此外,随着近年来复杂网络分析技术与方法的发展(如网络社区探测、聚类算法和网络表示学习等),以及科学文献数据的爆发式增长,可知关于科学知识网络中社区(社区、社群、群落)发展过程中的扩张、收敛时序演变问题,在研究理论、研究数据和研究技术方法 3 个方面都具备展开研究的基础。

综上所述,笔者在 T. S. Kuhn 的科学发展范式理论与科学共同体学说、D. J. Price 的科学知识增长定量研究基础上,综合 D. J. Price 关于无形学院的相关研究,以复杂网络思维为指导,从科学知识扩散角度切入,探索科学知识扩散过程中的群聚性问题和科学知识扩散过程中的群聚性问题,对科学知识网络中社区(社区、社群、群落)发展过程中的扩张、收敛时序演变过程进行动态跟踪建模,从而揭示科学知识网络时序变化过程中社区扩张与收敛的基本模式与规律,以期为知识扩散相关研究提供有益的研究视角。

3 研究设计

3.1 研究数据

笔者以 PubMed 数据库所收录的“医疗健康信息”领域的文献数据为研究数据,具体检索策略是选择 PubMed 数据库为检索数据库,以“medical/health informatics/information”为检索词进行题名检索,发文年代范围不限,具体检索式为(((medical information[Title]) OR health information[Title]) OR health informatics[Title]) OR medical informatics[Title] AND “NIH grants”[Filter]。

以检索结果为领域核心文献,导出 PMID,然后基

于 R 编程进行引文爬取构建引文网络,共得到 1981 年 –2019 年间发表的 5 643 篇文献,然后以这些文献的 PMID 号(在线检索工具:https://www.ncbi.nlm.nih.gov/sites/batchentrez)检索下载相应题录数据保存至本地以备后续研究。由于研究数据时间跨度较大,为了有效揭示知识扩散过程中的社区扩张与收敛模式,所以对其进行时间段划分,具体将 1981 年 –2019 年划分为 8 个时间段(对应 8 个时期),由于研究数据时间跨度近 40 年,并且在早期 1981 年 –2000 年文献数量较少,如果严格按照自然年份等距划分时期会导致前几个时期与后几个时期相差过于悬殊不利于后续模型的构建,因此笔者灵活划分时间窗口,以保证各个时期的文献数量不会相差过大,并且符合科学文献增长规律(即普莱斯文献指数增长规律),从而保证后续研究的科学性、有效性。各个时期文献数量分布情况如图 1 所示:

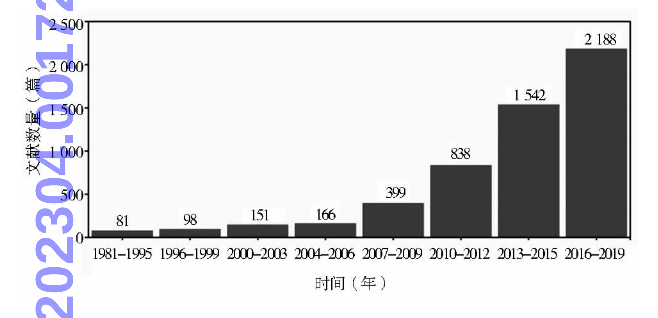


图 1 文献数量时期分布

3.2 研究方法与流程

本研究的基本目标是基于文献引用数据对科学知识扩散中的社区扩张与收敛时序变化现象进行客观测度和分析,并尝试通过数学模型总结归纳出社区扩张与收敛时序变化现象后面隐含的固有模式与规律。

围绕研究目标,提出研究方法与基本流程:首先基于文献引用关系建立矩阵进而构建学科知识网络,采用复杂网络分析中的 Louvain 社区探测算法对领域知识网络进行划分,然后利用网络表示学习技术(graph embedding)进行社区扩张与收敛特征表示与计算,并以时间序列为逻辑线索,对科学知识网络中社区(社群、群落)的扩张、收敛演变过程进行动态跟踪建模,从而揭示科学知识网络时序扩散变化过程中社区扩张与收敛的基本模式与规律,下面对主要研究方法流程进行分析。

3.2.1 科学知识网络构建

笔者将科学知识网络定义为描述科学文献(节点)及其引用关系(边)按照时间顺序相互作用的数据结构模型。从可视化层面讲,即节点(文献)及其边(引用关系)随着时间推移不断变化构成的图 Gt =(Vt, Et),表示在任意时间段[0, n](0 ≦ t ≦ n)中科学知识网络集合。

科学知识网络构建是进行后续研究的基础,为了实现科学知识网络社区扩张与收敛模式分析,需要实现对不同时间窗口的科学知识网络进行动态分析,本文中科学知识网络构建可以概括为以下两个子步骤:

(1)科学知识网络是一个增长型网络,为了对科学知识网络中社区(社群、群落)的扩张、收敛演变过程进行动态跟踪建模,因此需要按照年度划分时间段 T = {t1 – t2, t2 – t3, ……, tn – tm}, n < m。

(2)根据上一步划分的时间段 T = {t1 – t2, t2 – t3, ……, tn – tm},分别基于引用关系构建文献引用矩阵,进而建立不同时期的科学知识网络 $G_T = \{G_1, G_2, G_3 \cdots\cdots, G_n\}$,为后续研究奠定数据基础。

不同时期的科学知识网络构建采用时间切片的方法(也可称半累积科学知识网络[24]),是指在所分析的时间段中引文网络由两部分组成:①该时间段内发表的文献(施引文献);②这些文献所引用的本时间段以及之前所有时间的文献(被引文献)。如果分析全累计的引文网络,则过去引文信息的积累很容易湮没当前文献所揭示的社区,所以,笔者通过构建不同时间段的引文网络切片以有效探索、揭示知识扩散过程中社区时序演变情况。

3.2.2 基于 Louvain 算法的社区探测

笔者利用 Louvain 算法[25-26]进行社区探测,Louvain 算法是基于模块度(modularity)的社区发现算法,可以有效探测层次性社区结构,模块度是目前评价社区探测结果的主要指标,模块度越大意味着社区发现的效果越好。

模块度这一概念由 M. E. Newman 等[27](2004)首次提出,并受到了复杂网络领域学者的关于与认可,以之为基础提出了众多社区探测算法,其中 Louvain 算法是代表成果之一。Louvain 算法在效率和效果上都表现比较好,主要有易于理解、非监督和计算快速等优点,其中,目前 Python 环境下的 NetworkX 工具包和复杂网络分析软件 Gephi 中都集成了 Louvain 算法功能。具体利用 Louvain 算法识别每个时间窗口内的知识社区,可以细分为原始划分、模块度优化、社区聚合和社区探测 4 个步骤。其中,关键步骤为模块度优化,模块度计算方法如公式(1)[25]所示:

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_i, c_j) \qquad 公式(1)$$

其中，Q 代表模块度；A_{ij} 表示复杂网络中 i 和 j 之间边的权重；k_i，k_j 表示与节点 i 和 j 相连的所有边的权重之和；c_i，c_j 表示节点 i 和 j 的所属社区，当 i 和 j 属于同一社区 $\delta(c_i, c)=1$，不属于同一社区时等于 0；Q 的最大值为 1，Q 越接近这个值，就说明社区结构越明显。

3.2.3 基于 Graph Embedding 的社区扩张与收敛特征表示与计算

传统经典社会网络分析方法（度、中心性和密度等）无法有效测度引文网络中社区的扩张与收敛特征，随着深度学习技术的快速发展，网络表示学习法（Graph Embedding Method，GEM），也称图嵌入（将节点表示为实数值向量，同时保持网络结构和节点固有属性的方法，利用深度学习算法自动提取每个网络节点的特征，并形成 embedding 嵌入），是从 Word2vec 等[28－29]发展而来的 Embedding 技术的最新延伸，近年来在推荐系统、计算广告领域应用广泛，其中，Deepwalk、LINE、Node2vec 等是比较有代表性的方法。

笔者尝试利用网络表示学习技术以更好地测度社区扩张与收敛特征，研究中使用斯坦福大学开源的 Node2vec 进行社区扩张与收敛特征表示与计算。Node2vec 是将网络中的节点表征为实数值向量的算法模型，其利用深度学习的思想，可以通过一个三层神经网络（输入层－隐藏层－输出层）把每个节点映射成 K 维实数向量，将网络中任意两个节点的相关关系，转换为对应两个向量的相关关系，利于计算存储，不用再手动提特征（自适应性），其中，Node2vec 通过定义一个目标函数 f(u)来表示学习节点的局部邻居结构，可以将属于相同社区以及结构相似的节点学习得到相近特征，该目标函数如公式(2)[30]所示：

$$\begin{cases} f(u) = max_f \sum_{u \in V} \log P(N_S(u) | f(u)) \\ P(N_S(u) | f(u)) = \prod_{n_i \in N_S(u)} P(n_i | f(u)) \end{cases} \qquad 公式(2)$$

其中，$f(u)$ 为将节点 u 映射为 embedding 向量的目标函数；V 表示网络中节点的集合，S 指得到节点邻居 N 的策略，$N_S(u)$ 表示通过采样策略 S 采样出的节点 u 的近邻顶点集合，Node2vec 模型算法[30]如图 2 所示：

```
Algorithm 1 The node2vec algorithm.
LearnFeatures (Graph G = (V, E, W), Dimensions d, Walks per
    node r, Walk length l, Context size k, Return p, In-out q)
    π = PreprocessModifiedWeights(G, p, q)
    G' = (V, E, π)
    Initialize walks to Empty
    for iter = 1 to r do
        for all nodes u ∈ V do
            walk = node2vecWalk(G', u, l)
            Append walk to walks
    f = StochasticGradientDescent(k, d, walks)
    return f

node2vecWalk (Graph G' = (V, E, π), Start node u, Length l)
    Inititalize walk to [u]
    for walk_iter = 1 to l do
        curr = walk[−1]
        V_curr = GetNeighbors(curr, G')
        s = AliasSample(V_curr, π)
        Append s to walk
    return walk
```

图 2　Node2vec 模型算法

本研究主要利用 Python 语言进行社区扩张与收敛特征表示与计算，在具体处理步骤中，首先基于 Node2vec 模型算法（https://github.com/aditya-grover/node2vec）将引文网络中的每个节点表示成可计算的 K 维向量，然后结合上一步中的社区划分结果，计算各个社区节点（向量）之间的距离进而可得社区所占区域面积的大小（以各个社区内节点最大距离为区域直径），然后通过分析各个社区在不同时间段的区域面积时序变化情况，从而表征与测度引文网络时序变化过程中社区的扩张与收敛情况，基本思路可归纳概括为如图 3 所示：

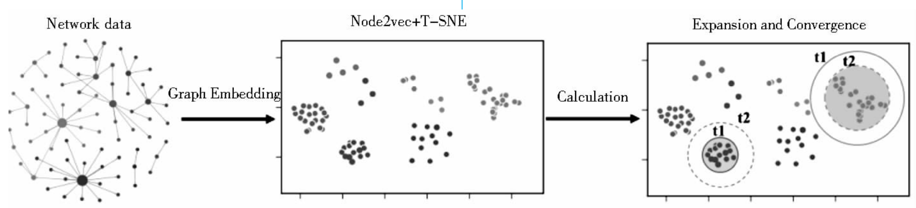


图 3　基于 Graph Embedding 的社区扩张与收敛特征计算基本思路

3.2.4 社区扩张与收敛模型构建与模式分析

最后，以时间序列（划分的时间窗口）为逻辑线索，对科学知识网络中社区（社群、群落）的扩张、收敛演变过程进行动态跟踪、拟合建模，从而对科学知识网络时序扩散变化过程中社区扩张与收敛现象进行分析，并尝试归纳其基本模式，具体可以分为 2 个子步骤：

（1）在对社区扩张与收敛时序数据进行拟合之

前，首先根据上一步计算得到的社区扩张与收敛特征计算结果的时序数据绘制变化趋势图（折线图、散点图等）；

（2）然后，通过观察时序数据可视化结果来分析社区时序变化特征，以确定应该使用何种函数曲线进行拟合，即确定拟合的函数模型类别。

目前典型的拟合函数模型主要有指数函数、幂函数、双曲线函数、指数函数和 S 形曲线函数等，基本公式与函数图像如图 4 所示：

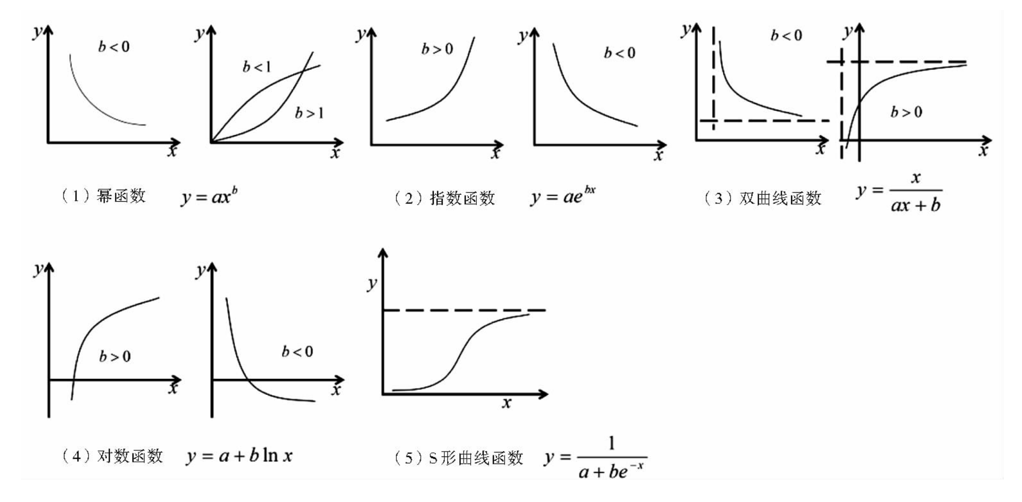


图 4　目前典型的拟合函数模型示意

通过上述方法步骤的处理分析，对科学知识扩散中的社区扩张与收敛时序变化现象进行了客观测度和分析，即基于科学文献之间的引用关系揭示了中观视角下（社区）的知识扩散规律，并通过科学的数学模型总结归纳出社区扩张与收敛时序变化现象背后隐含的固有模式，对于拓展、深化科学知识网络的扩散、传递相关研究具有一定的理论与实践意义。

4　结果分析

4.1　科学知识网络构建与社区探测

根据研究方法中提出的流程步骤，分别构建“医疗健康信息（health informatics）”领域 8 个时期的初始引文网络；然后，利用 Louvian 算法对初始引文网络进行社区探测操作，对输入的文献引用数据进行模块化分析，将初始模块分割参数（resolution）设置为默认值 5，并根据实际划分结果对各个时期的 resolution 值进行微调以提高社区探测结果准确性，最终获得各个时期的社区探测结果，各个时期的引文网络社区探测结果如图 5 所示，节点大小正比于中心性，布局方式（layout）采用 FR（Fruchterman Reingold）布局。

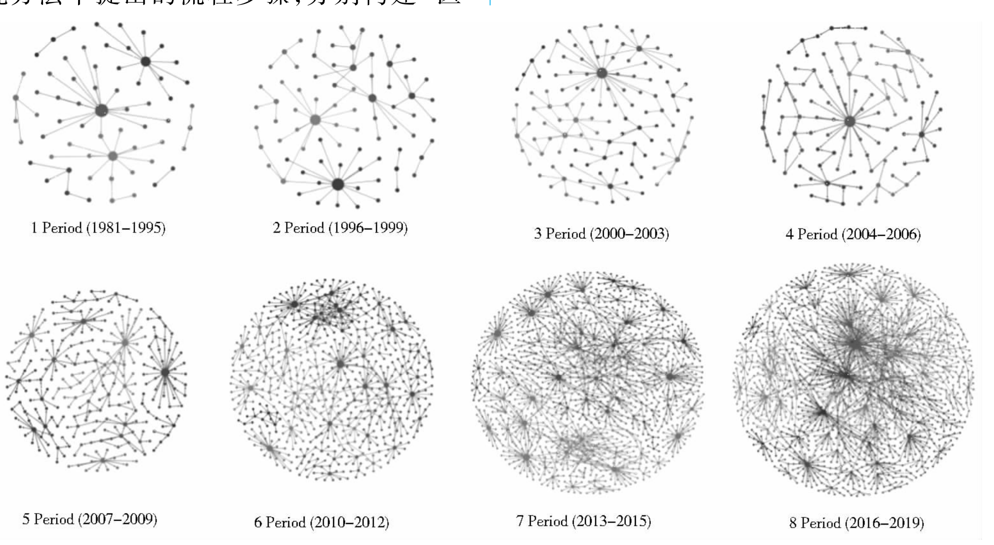


图 5　各个时期社区探测结果

图5表示8个时期的初始引文网络,展现了各个时期初始引文网络的社区探测结果,即通过文献节点的联系和聚集紧密程度来揭示社区结构。分析图5可知,随着时间的推移,医疗健康信息领域初始引文网络不断生长,并且各个时期引文网络都呈现出清晰的社区结构,在引文网络社区时序演变过程中表现出一定的社区规模的扩张与收敛现象,下面继续对其进行进一步的揭示分析。

4.2 基于Node2vec的社区扩张与收敛特征计算结果

在上一步社区探测结果基础上,结合引文网络数据,利用Node2vec模型算法对各个时期引文网络进行计算从而提取每个网络节点的特征,将节点表示为128维(dimension)的实数值向量,从而得到各个时期社区节点向量(见表2),表中社区ID指各个节点所属社区,PMID表示各个节点的PubMed数据库中的唯一标识号,dimn表示节点的n维实数向量。

表2 各个时期社区节点向量(部分结果)

社区ID	PMID	dim1	dim2	dim3	dim4	dim5	dim6	dim7	……	Dim128
社区4	1 482 884	0.364 5	-0.182 0	-0.045 9	0.173 1	0.468 9	0.090 2	0.164 3	……	-0.182 5
社区4	1 482 926	0.336 6	-0.261 7	-0.093 4	0.186 0	0.416 4	0.120 1	0.100 6	……	-0.081 2
社区5	1 482 948	0.010 4	-0.139 5	-0.064 6	-0.227 2	0.383 4	0.024 4	-0.215 2	……	0.218 2
社区4	1 482 993	0.371 9	-0.220 2	-0.078 8	0.198 7	0.494 0	0.099 8	0.139 2	……	-0.147 8
社区3	1 537 018	0.221 3	-0.445 8	0.070 1	0.276 4	0.470 5	0.535 8	-0.035 3	……	0.132 2
社区4	1 582 194	0.378 9	-0.196 6	-0.027 1	0.190 6	0.558 5	0.096 2	0.186 9	……	-0.191 9
社区5	1 807 576	-0.002 6	-0.142 3	-0.076 2	-0.244 4	0.398 1	0.019 9	-0.234 6	……	0.239 1
社区6	1 807 625	0.270 3	-0.087 3	0.157 7	0.023 3	0.201 8	0.063 6	0.238 5	……	-0.305 5
社区5	1 807 701	0.006 9	-0.151 7	-0.072 5	-0.246 7	0.408 2	0.024 7	-0.235 6	……	0.244 9
社区5	1 807 737	0.002 2	-0.139 4	-0.069 6	-0.243 4	0.372 5	0.017 3	-0.227 7	……	0.224 5
社区5	1 807 738	0.004 5	-0.144 0	-0.066 3	-0.226 8	0.390 2	0.023 2	-0.223 3	……	0.230 2

然后将各个时期的社区节点向量用T-SNE[31](T-Distribution Stochastic Neighbour Embedding)映射到二维平面进行可视化分析(见图6)。T-SNE(T分布随机近邻嵌入)是一种用于降维的机器学习方法,能帮助识别相关联的模式,主要的优势是保持高维数据局部结构的能力,即高维数据空间中距离相近的点投影到低维中仍然相近。

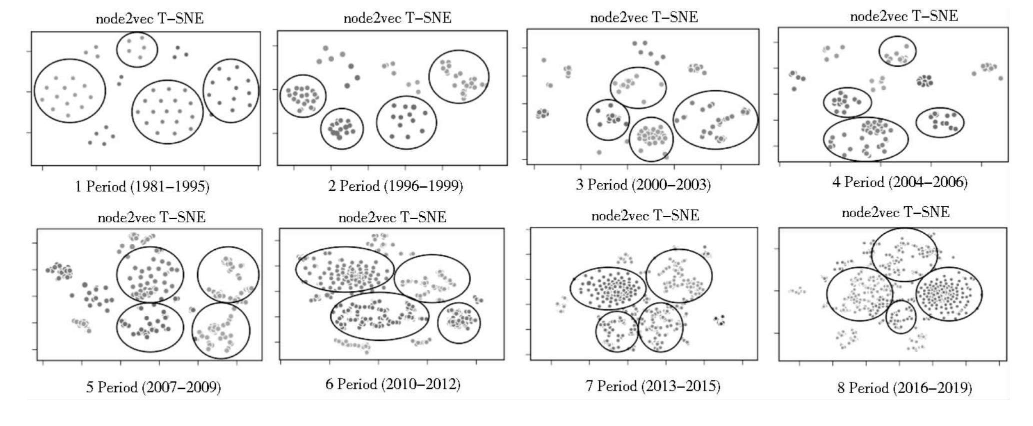


图6 各个时期社区节点特征提取及其T-SNE可视化

将图6和图5进行对比分析,图6中反映的社区节点聚集和分布情况,与传统经典社区探测结果基本一致,由此可知,基于Node2vec对社区节点进行特征提取并表示为向量,能够有效揭示各个时期引文网络的社区结构,在一定程度上说明,通过计算社区节点向量来测度社区规模的大小是可行有效的。因此,利用各个时期社区节点向量结果,分别计算社区节点之间的距离(具体使用欧几里得度量euclidean metric,也称欧氏距离指在n维空间中两个点之间的真实距离),选取社区节点间的最大距离,以之为直径计算各个社区区域面积大小,进而可以对不同时段社区所占区域面积大小进行时序数据构建,然后在后续步骤中通过进行数学建模以有效测度社区规模的扩张与收敛。

4.3 医疗健康信息领域社区扩张与收敛时序分析

笔者结合全时期网络节点向量T-SNE可视化和社区规模时序变化数据(见图7),对医疗健康信息领域

整体规模 TOP5 的热点社区(互联网健康信息、健康信息行为、电子健康信息系统与技术、健康信息评估和健康管理)进行社区规模变化趋势解读分析,并为下一步的社区扩张与收敛模型构建及模式分析奠定基础。

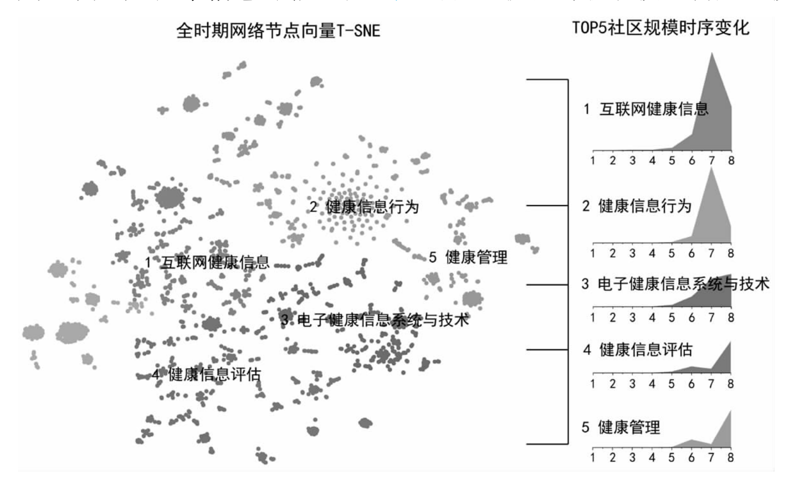


图 7 全时期网络节点向量 T-SNE 可视化与 TOP5 社区时序变化

(1)互联网健康信息。互联网技术的快速发展促使健康信息的获取更加方便快捷,公众获取健康信息的手段不仅仅依赖于医院、医生,而是逐步转向信息获取方便快捷的互联网平台,在此背景下促进了健康服务平台的产生和发展;公众对于健康信息的可访问性的影响因素知之甚少,存在着信息质量差、使用一级搜索引擎和简单搜索术语访问健康信息效率不高的问题,如何实现健康信息可访问性、提高检索效率也是研究的重中之重;此外,大数据环境下公众的健康信息安全、隐私保护等问题也是目前研究的重点内容。根据图 7 可以看出,该社区在前 5 个阶段社区规模呈现平稳的发展趋势,第 6 阶段快速上升到峰值,第 7、8 阶段呈现下降趋势,该趋势反映互联网健康信息为健康信息领域的重点及热点内容。

(2)健康信息行为。公众健康信息需要的不断增长促使公众获取健康信息的途径逐渐多样化,互联网技术及移动互联网技术的发展更为健康信息的获取提供了更加方便快捷有效的方式。公众对健康信息的持续关注促使了一系列健康信息行为的发生,信息检索、信息获取等信息行为以及信息行为的影响因素逐渐成为研究重点。根据图 7 可以看出,该社区规模时序变化与互联网健康信息基本一致,都呈现出平稳发展到快速发展再到逐渐下降的趋势,是健康信息领域的重点及热点内容。

(3)电子健康信息系统与技术。目前,国外对于电子健康信息系统与技术的建设相对成熟,既有关于电子健康信息系统完善的构建流程的概念模型,促进健康信息技术的不断完善和发展,也有能够用于实践的完善的平台和系统。根据图 7 可以看出,该社区在第 6 时期前社区规模时序变化呈现平稳的发展趋势并在第 6 时期后逐步上升,反映该社区逐渐成为医疗健康信息领域的重点研究内容。

(4)健康信息评估。在网络环境下公众能够及时便利地获取相关健康信息,但网上资源质量参差不齐,如何保证健康信息的质量、实现质量控制逐渐成为研究热点,因此健康信息评估的相关研究逐渐增多。关于健康信息评估目前主要集中在影响健康信息的因素与评估体系的构建,未来研究更加侧重于技术手段实现信息评价。根据图 7 可以看出,该社区的社区规模时序变化在第 7 阶段前呈现平稳趋势,第 7 时期后快速发展,反映该社区将在未来几年内成为健康信息领域的研究热点。

(5)健康管理。随着近年来人们生活方式和健康理念的转变,健康管理在新的健康模式下也表现出了新的特点,逐渐成为一种新兴的健康服务理念和服务方式。在此背景下,健康管理服务的发展带动了学者对健康管理服务系统的研究,目前主要集中于以下 3 个方面:①以各类疾病的治疗为核心的医疗健康管理服务系统;②健康管理服务技术;③健康管理系统或体系的构成。根据图 7 可以看出,该社区规模时序变化的发展与健康信息相似,同样具有成为领域研究热点的潜力。

4.4 科学知识网络社区扩张与收敛模型构建及模式分析

在上述分析的基础上,对社区规模时序变化数据进行拟合建模。目前典型的拟合函数模型主要有指数函数、幂函数、双曲线函数、指数函数和 S 形曲线函数

岳丽欣，周晓英，刘自强.科学知识网络扩散中的社区扩张与收敛模式特征分析——以医疗健康信息领域为例[J].图书情报工作,2020,64(14):63－73.

等,由观测研究数据(绘制折线图,观测变化趋势)和普赖斯曲线(文献增长规律)可知,科学知识网络中社区规模时序变化过程更加契合 S 形曲线函数(其中代表性函数 Logistic 模型),因此,笔者利用 S 形曲线函数对社区扩张与收敛时序变化数据进行 S 形曲线拟合以构建数学模型,结果如图 8 所示,图中展示了互联网健康信息、健康信息行为和电子健康信息系统与技术等 5 个热点社区(整体规模 TOP5)模型构建结果。

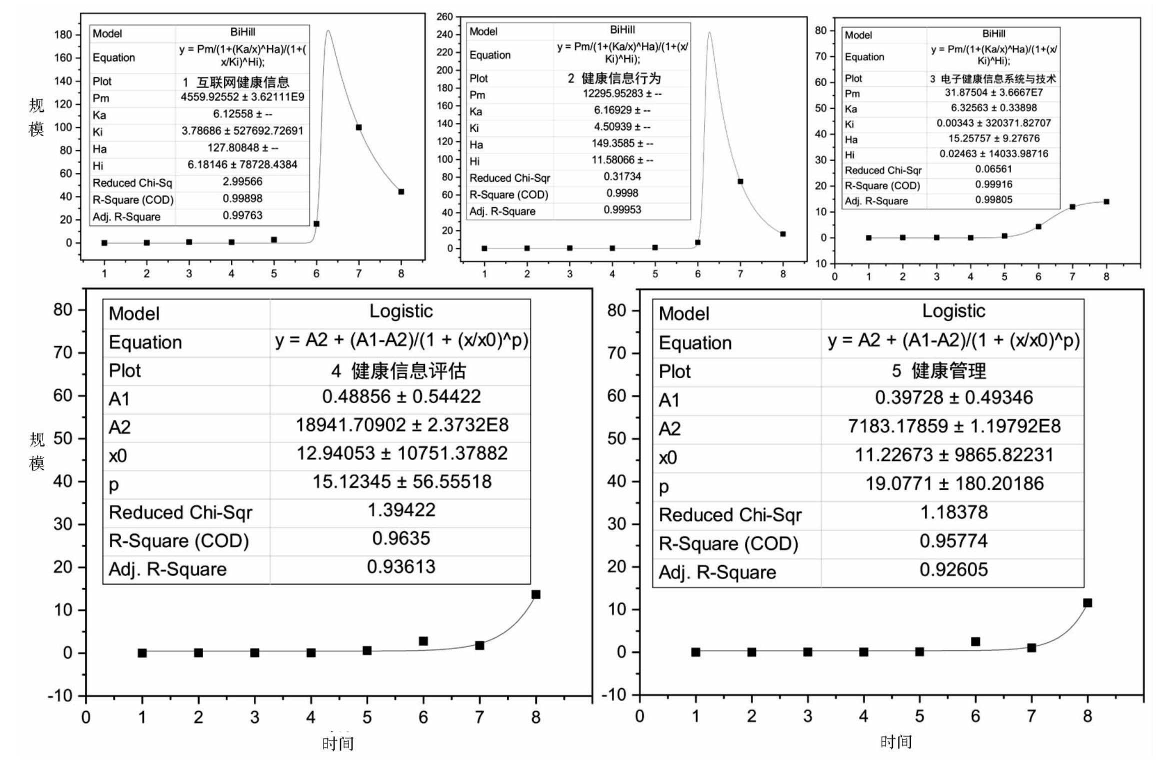


图 8　TOP5 热点社区扩张与收敛模型构建

由图 8 可知,校正决定系数(Adj. R-Square)都在 0.9 以上,说明 5 个热点社区扩张与收敛模型构建结果良好,校正决定系数可以反映模型拟合结果的好坏(越接近 1,说明拟合结果越好,负数说明结果偏差太大)。

结合图 8 医疗健康信息领域 Top5 社区扩张与收敛模型,对互联网健康信息、健康信息行为和电子健康信息系统与技术等社区的扩张与收敛模式及其基本特征进行归纳总结分析。如表 3 所示:

表 3　社区扩张与收敛模式及其基本特征

社区模式	模式特征	函数类型	数理模型	基本公式
扩张模式	科学范式涌现; 社区成员少; 社区成员快速增加; 处于成长阶段	S 形曲线函数	Logistic 模型	$y=\frac{A_1-A_2}{1+(x/x_0)^p}+A_2$
收敛模式	科学范式聚焦、深入; 社区群体明显; 社区成员逐渐减少; 处于衰退阶段	S 形曲线函数	BiHill 模型	$y=\frac{P_m}{\left[1+\left(\frac{K_a}{x}\right)^{H_a}\right]\left[1+\left(\frac{x}{K_i}\right)^{H_i}\right]}$

分析可知,随着近几十年信息技术的不断进步,医疗健康信息领域发展良好,目前该领域内社区扩张与收敛模式共存;其中,互联网健康信息、健康信息行为和电子健康信息系统与技术作为领域研究核心,目前社区特征明显社区成员较多。总体来看,处于社区扩张模式向社区收敛模式转变的时期,发展趋势符合 S 形曲线函数中的 BiHill 模型(已过最大值,处于递减阶段,对应社区收敛模式);健康信息评估和健康管理作为近年来的新兴社区,成长态势明显,社区成员不断增加,社区规模不断扩大,发展趋势符合 S 形曲线函数中的 Logistic 模型(快速成长期,对应社区扩张模式)。

经过上述分析可得出以下结论,科学知识网络扩

散过程中的社区扩张与收敛时序变化符合某些函数模型，可以描述社区扩张和收敛时序变化模式特征，具体如下：

(1)社区扩张模式的基本特征为科学范式涌现，社区成员少，社区成员快速增加，处于科学生命周期中的快速成长阶段，符合S形曲线函数中的Logistic模型，即快速成长期，对应社区扩张模式；

(2)社区收敛模式的基本特征为该模式下的社区内部研究的科学范式聚焦、深入，社区群体明显，社区成员逐渐减少，处于科学生命周期中的衰退阶段，符合S形曲线函数中的BiHill模型，即已过最大值，处于成员递减阶段，对应社区收敛模式。

由上述科学知识网络扩散过程中的社区扩张与收敛模式与特征分析结果可知，本研究一方面印证了科学知识增长机制(文献指数增长规律)，即学科知识增长、知识扩散过程中的集群(社区)在前期发展趋势符合指数增长特征，但是也暴露出文献指数增长规律不完备的一面：未考虑学科领域知识单元间的关联关系(社区)及其互动规律，即文献指数增长规律无法准确、有效地描述知识集群、知识社区层面的知识增长机制。而本研究探索了文献指数增长规律在科学知识扩散、知识网络社区演化等先进理论与方法视角下的新特征、新模式，有助于揭示知识发展演化过程中交叉、衍生、融合等现象背后的规律；并且本研究通过定量数学建模科学地揭示了知识在科学共同体中的扩散的时序变化模式与特征，将科学知识扩散、知识增长机制、知识网络社区演化、知识在科学共同体中的扩散等研究进行了有机融合，分析了科学知识扩散过程中的群聚性问题，对科学知识网络中社区(社区、社群、群落)发展过程中的扩张、收敛时序演变过程进行动态跟踪建模，从而揭示科学知识网络时序变化过程中社区扩张与收敛的基本模式与特征，为目前科学知识增长、知识扩散和知识社区演化等相关研究提供了有益的研究视角，具有一定的借鉴和参考意义。

5 结语

笔者以科学知识扩散过程中的社区扩张与收敛模式为研究问题，以引文网络为研究数据，在经典复杂网络分析理论与技术基础上，利用深度学习技术结合时序分析方法，对科学知识网络中社区的扩张、收敛演变过程进行动态跟踪与建模，从而揭示科学知识扩散过程中的社区扩张与收敛基本模式与规律。本研究以医疗健康信息领域进行了案例研究，研究结果表明：社区扩张模式的发展趋势符合S形曲线函数中的Logistic模型，社区收敛模式的发展趋势符合S形曲线函数中的BiHill模型。本文还存在一些局限与不足，仅以医疗健康信息领域为例进行了研究可能造成结果不准确，未深入到具体文本内容进行科学知识扩散过程中的社区扩张与收敛模式研究，此外，如果采用的社区划分算法方法可能会导致结论存在一定差异。接下来的工作，将进一步扩大研究数据，数据应当涉及理科、工科以及人文社科等其他不同的学科领域，并尝试研究具体文本内容维度下的研究主题的扩张与收敛问题。

参考文献：

[1] 邱均平，李小涛. 基于引文网络挖掘和时序分析的知识扩散研究[J]. 情报理论与实践，2014(7):5-10.

[2] 李纲，巴志超. 科研合作超网络下的知识扩散演化模型研究[J]. 情报学报，2017(3):58-68.

[3] 岳增慧，许海云. 学科引证网络知识扩散特征研究[J]. 情报学报，2019，38(1):5-16.

[4] NOYONS E C M，RAAN A F J V. Monitoring scientific developments from a dynamic perspective：self-organized structuring to map neural network research[J]. Journal of the Association for Information Science & Technology，1998，49(1):68-81.

[5] 雷迭斯多夫. 科学计量学的挑战：科学交流的发展、测度和自组织[M]. 乌云，译. 北京：科学技术文献出版社，2003.

[6] LEYDESDORFF L，COZZENS S，PETER V D B. Tracking areas of strategic importance using scientometric journal mappings[J]. Research policy，1994，23(2):217-229.

[7] LEYDESDORFF L. Statistics for the dynamic analysis of scientometric data：the evolution of the sciences in terms of trajectories and regimes[J]. Scientometrics，2013，96(3):731-741.

[8] POPPER K R. The logic of scientific discovery[J]. Yinshan academic journal，2005，12(11)：53-54.

[9] 靖继鹏，马费成，张向. 情报科学理论[M]. 北京：科学出版社，2009.

[10] 刘则渊. 跨越学术分水岭[M]. 北京：人民出版社，2012.

[11] 万昊. 科学知识规模增长模式研究－基于数学建模和仿真论证[D]. 北京：中国科学院大学，2017.

[12] 刘自强，许海云，罗瑞，等. 基于主题关联分析的科技互动模式识别方法研究[J]. 情报学报，2019，38(10)：997-1011.

[13] 安宁，滕广青，白淑春，等. 领域知识聚类性的动态演化分析[J]. 图书情报工作，2018，62(10):85-93.

[14] GIRVAN M，NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences of the United States of America，2002，99(12)：7821-7826.

[15] BETTENCOURT L M A，KAISER D I，KAUR J. Scientific discovery and topological transitions in collaboration networks[J]. Journal of informetrics，2009，3(3):210-221.

[16] 白如江，冷伏海. k-clique社区知识创新演化方法研究[J]. 图书

情报工作,2013,57(17):94 －99.
[17] 王晓光,程齐凯. 基于 NEViewer 的学科主题演化可视化分析[J]. 情报学报, 2013, 32(9):900 －911.
[18] 滕广青. 基于频度演化的领域知识关联关系涌现[J]. 中国图书馆学报,2018,44(3):79 －95.
[19] 滕广青. 关联驱动的领域知识群落生长[J]. 中国图书馆学报, 2017,43(3):58 －71.
[20] KUHN T S. The structure of scientific revolutions[M]. Chicago: University of Chicago Press, 1962.
[21] PRICE D J. Science since babylon[M]. New Haven: Yale University Press, 1961.
[22] PRICE D J. Little science, big science[M]. New York: Columbia University Press, 1963.
[23] DIANA C. Invisible colleges-diffusion of knowledge in scientific communities[M]. Chicago: The University of Chicago Press, 1972.
[24] 罗双玲,张文琪,夏昊翔. 基于半积累引文网络社区发现的学科领域主题演化分析——以"合作演化"领域为例[J]. 情报学报,2017,36(1):100 －110.
[25] BLONDEL V D,GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks[J]. Journal of statistical mechanics: theory and experiment, 2008, 2008(10): 10008.
[26] BLONDEL V D. Louvain algorithm[EB/OL]. [2019 －07 －17]. https://perso.uclouvain.be/vincent.blondel/research/louvain.html.
[27] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical review, 2004, 69(2):108 －113.
[28] 周练. Word2vec 的工作原理及应用探究[J]. 科技情报开发与经济,2015,25(2):145 －148.
[29] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Advances in neural information processing systems 26. Cambridge: Neural Information Processing Systems Foundation Inc., 2013: 3111 －3119.
[30] GROVER A, LESKOVEC J. Node2vec: scalable feature learning for networks[C]// Proceeding of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2016:855 －864.
[31] LAURENS V D M, HINTON G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11):2579 －2605.

作者贡献说明：
岳丽欣：论文思路与框架设计、撰写与修改；
周晓英：提出论文思路，论文定稿与修改；
刘自强：论文框架设计、实验设计以及论文修改。

Analysis on the Characteristics of Community Expansion and Convergence Mode in the Diffusion of Scientific Knowledge Network——Take the Field of Medical Health Information as an Example

Yue Lixin[1] Zhou Xiaoying[1] Liu Ziqiang[2,3]
[1] School of Information Resources Management, Renmin University of China, Beijing 100872
[2] Chengdu Library of Chinese Academy of Sciences, Chengdu 610041
[3] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] Knowledge units in scientific knowledge networks show certain clustering and communality, revealing the basic patterns and rules of community expansion and convergence in the process of changing the time series of scientific knowledge networks, which has certain significance for expanding and deepening the research on the diffusion and transmission of scientific knowledge. [**Method/process**] Firstly, the adjacency matrix was built based on the citation relation, and then the subject knowledge network was constructed. The Louvain community detection algorithm in complex network analysis is used to divide the domain knowledge network into communities. Then, the Graph Embedding technique was used to represent and calculate the community expansion and convergence characteristics. Finally, the time series was used as the time series. Logical clues were used to dynamically track and model the process of expansion and convergence of different communities, so as to reveal the basic patterns and laws of community expansion and convergence in the process of time series change of scientific knowledge network. [**Result/conclusion**] A case study in the field of health information shows that the trend of community expansion conforms to the Logistic model in the S-shaped curve function and the trend of community convergence conforms to the BiHill model in the S-shaped curve function.

Keywords: knowledge network community detection graph embedding expansion model convergence model